CAUSAL EPINETS: SCALABLE INFERENCE FOR BOUNDS ON THE PROBABILITY OF POSITIVE INDIVIDUAL TREATMENT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does a treatment help this individual? For binary treatments and outcomes, the Probability of Necessity and Sufficiency (PNS) formalises this question as the probability of a positive Individual Treatment Effect. In general, PNS is not point identified, but prior work shows that sharp bounds can be obtained by combining randomised and observational data without assuming observational unconfoundedness or positivity. These identification results characterise what can be learned given exact knowledge of the underlying distributions; in practice, however, PNS bounds must be estimated from finite samples. We show that naive plug-in estimation of PNS bounds suffers from two distinct finite-sample pathologies and develop a principled, scalable framework that translates the identification theory into practical estimation with high-dimensional covariates while explicitly addressing both.

1 INTRODUCTION

Personalised decision-making relies on understanding how a specific action or treatment affects a specific individual. To determine this, we must measure the causal effect of the action, which is the difference between an individual's outcome if the action is taken versus if it is not, at the exact same moment in time. In causal inference, the Individual Treatment Effect (ITE) captures this difference but is inherently unobservable, since only one of these two potential outcomes can be observed for the same individual. Much of causal inference therefore focuses on the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE), which average individual effects across a population or among individuals with the same observed covariates. Standard approaches to estimating these quantities from observational data rely on untestable assumptions. Even when identified, however, average effects generally do not determine the probability that an individual benefits from treatment: different mixtures of beneficial and harmful responses can produce the same average effect. This distinction motivates a complementary target for personalised decision-making: the probability of benefit.

For binary outcomes, the *Probability of Necessity and Sufficiency* (PNS), introduced by Pearl (1999), is the joint counterfactual probability $P(Y_1 = 1, Y_0 = 0)$: the probability that the outcome would occur under treatment but not under control. When the transition from 0 under control to 1 under treatment represents a beneficial response, Mueller & Pearl (2023) refer to PNS as the probability of benefit. Although PNS is generally not point identified, Tian & Pearl (2000) show that it can be sharply bounded by combining experimental and observational data. Experimental data identify interventional response probabilities, while observational data constrain the joint distribution of treatment and outcome. Together, these sources restrict the set of admissible structural causal models. Importantly, the resulting bounds do not require unconfoundedness or positivity of the observational treatment assignment.

These identification results characterise what can be learned given exact knowledge of the relevant probabilities. In practice, however, PNS bounds must be estimated from finite samples, often conditional on high-dimensional covariates. A natural approach is to estimate each probability from data and substitute the estimates into the bound expressions. Such naive plug-in estimators face two fundamental difficulties. First, the estimated probabilities need not be jointly compatible with any

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

underlying structural causal model. Second, PNS bounds are defined through maxima and minima over multiple estimated terms. This intersection-bound structure can bias estimated lower bounds upward and upper bounds downward, making the bounds appear more informative than the data warrant (Chernozhukov et al., 2013b). Reliable estimation and inference must therefore address both structural compatibility and the effects of sampling uncertainty on the max–min operators.

We develop an end-to-end framework for estimating Tian–Pearl PNS bounds and quantifying their uncertainty from finite samples with high-dimensional covariates. Our approach integrates three components. First, we introduce an anchored, constrained multi-head neural architecture that jointly models observational and interventional quantities while guaranteeing compatibility with probability axioms and a shared structural causal model by construction. Second, we show how PNS bounds fit naturally into the intersection-bounds framework and adapt precision-corrected inference to obtain confidence intervals that account for the bias induced by max–min operators. Third, we develop a scalable uncertainty quantification procedure based on Epistemic Neural Networks (ENNs) (Osband et al., 2023), providing a practical surrogate for resampling-based inference and enabling precision correction in modern over-parameterised models. Together, these components connect population-level identification results to practical estimation and inference, providing a principled approach to studying the probability of benefit beyond average treatment effects.

The remainder of the paper is organised as follows. Section 2 introduces the problem setup and reviews population-level PNS bounds. Section 3 presents the proposed anchored architecture, uncertainty modelling, and precision-corrected inference procedure. Section 4 reports empirical results. Section 5 discusses the findings, limitations, and future directions, and concludes the paper. Related work, theoretical results, and further experimental details are provided in the appendix.

2 Problem Setup

We consider a binary treatment $X \in \{0,1\}$, a binary outcome $Y \in \{0,1\}$, and pre-treatment covariates $Z \in \mathcal{Z}$. We work within a structural causal model in which the potential outcomes Y_x and $Y_{x'}$ are well defined, corresponding to the outcomes that would be observed under the interventions $\mathrm{do}(X=1)$ and $\mathrm{do}(X=0)$, respectively.

For each covariate value z, we summarise the relevant conditional quantities as follows. The *interventional success probabilities* are

$$\mu_1(z) := P(Y=1 \mid \mathrm{do}(X=1), Z=z)\,, \qquad \mu_0(z) \;\; := P(Y=1 \mid \mathrm{do}(X=0), Z=z)\,, \quad (1)$$

and the *observational joint probabilities* are

$$p_{xy}(z) := P(X=x, Y=y \mid Z=z), \;\; x,y \in \{0,1\}, \quad (2)$$

so that $p_{10}(z) + p_{11}(z) = P(X=1 \mid Z=z)$ and $p_{01}(z) + p_{00}(z) = P(X=0 \mid Z=z)$.

2.1 Probability of necessity and sufficiency

Following Tian & Pearl (2000), the *probability of necessity and sufficiency* (PNS) of X for Y is defined as

$$\mathrm{PNS} := P(Y_x = 1,\; Y_{x'} = 0)\,. \quad (3)$$

This quantity represents the fraction of units for which the treatment is both necessary and sufficient for the outcome.

We define the covariate-conditional PNS as $\mathrm{PNS}(z) := P(Y_x = 1,\; Y_{x'} = 0 \mid Z = z)$ and view the marginal PNS as $\mathrm{PNS} = \mathbb{E}_Z[\mathrm{PNS}(Z)]$.

2.1.1 Structural assumptions and constraint system

We adopt the structural assumptions used in Tian & Pearl (2000), formulated conditionally on $Z = z$.

Assumption 2.1 (Experimental identification). In the experimental regime, treatment assignment is randomised, possibly conditional on Z, so that X is independent of the potential outcomes given Z. Under this protocol, the interventional probabilities are identified by

$$\mu_x(z) := P(Y=1 \mid \mathrm{do}(X=x), Z=z) \qquad = P^{\mathrm{exp}}(Y=1 \mid X=x, Z=z), \qquad x \in \{0,1\}. \quad (4)$$

Assumption 2.2 (Stability across regimes). Experimental and observational data share the same structural equation for Y and the same distribution of exogenous variables. The treatment assignment mechanism for X may differ across regimes: in the experimental regime, X is assigned at random, while in the observational regime X may depend on Z and unobserved factors. Under this shared structural causal model, the observational joint distribution $(X, Y) \mid Z$ restricts the set of admissible interventional probabilities $(\mu_0(z), \mu_1(z))$ through the compatibility constraints

$$p_{11}(z) \le \mu_1(z) \le 1 - p_{10}(z), \qquad p_{01}(z) \quad \le \mu_0(z) \le 1 - p_{00}(z). \tag{5}$$

which must hold for any values of $(\mu_0(z), \mu_1(z))$ consistent with the observational distribution.

In the binary setting, these constraints are also jointly sufficient. For units observed with $X = 1$, consistency fixes $Y_1 = Y$, while the unobserved Y_0 values can be chosen to match any $\mu_0(z) \in [p_{01}(z), 1 - p_{00}(z)]$. Similarly, for units observed with $X = 0$, consistency fixes $Y_0 = Y$, while the unobserved Y_1 values can be chosen to match any $\mu_1(z) \in [p_{11}(z), 1 - p_{10}(z)]$. These choices concern different treatment groups and can therefore be made simultaneously. Hence every tuple satisfying the constraints in Equation 5 is compatible with at least one joint distribution of $(X, Y_0, Y_1) \mid Z = z$, which is the feasibility construction underlying the sharp bounds below (Tian & Pearl, 2000, Eqns. (15)–(22)).

2.1.2 Conditional and marginal PNS bounds

For a fixed z, the conditional PNS is a linear functional of the latent probabilities. Tian and Pearl derive sharp bounds by solving the corresponding linear programs.

In our notation, the bounds are

$$\mathrm{PNS}_{\mathrm{L}}(z) := \max\Big\{0,\ \mu_1(z) - \mu_0(z),\ p_{11}(z) + p_{01}(z) - \mu_0(z),\ \mu_1(z) - p_{11}(z) - p_{01}(z)\Big\}, \tag{6a}$$

$$\mathrm{PNS}_{\mathrm{U}}(z) := \min\Big\{\mu_1(z),\ 1 - \mu_0(z),\ p_{11}(z) + p_{00}(z),\ \mu_1(z) - \mu_0(z) + p_{10}(z) + p_{01}(z)\Big\}. \tag{6b}$$

These satisfy $\mathrm{PNS}_{\mathrm{L}}(z) \le \mathrm{PNS}(z) \le \mathrm{PNS}_{\mathrm{U}}(z)$, and the interval is sharp. The marginal PNS satisfies $\mathrm{PNS} \in \big[\beta_{\mathrm{L}}, \beta_{\mathrm{U}}\big]$, where $\beta_{\mathrm{L}} = \mathbb{E}_Z[\mathrm{PNS}_{\mathrm{L}}(z)]$ and $\beta_{\mathrm{U}} = \mathbb{E}_Z[\mathrm{PNS}_{\mathrm{U}}(z)]$.

The same framework extends to bounds on the probability of harm $P(Y_{x'} = 1, Y_x = 0 \mid Z = z)$. In the binary setting, $\mathrm{PNS}(z)$ coincides with the probability that the individual treatment effect is positive. Since this quantity is not point-identified, the Tian–Pearl bounds provide the tightest possible interval compatible with the observed data.

3 Proposed Approach/Contribution

In this section we will describe the proposed approach to overcome the two limitations of using naive plug-in estimators for fitting the experimental and observational distributions.

3.1 Ensuring estimates satisfy probability axioms

The structural constraints (Eq. 5) impose algebraic relationships between the observational joint probabilities $p_{xy}(z)$ and the interventional probabilities $\mu_x(z)$. Independently estimated models for $p_{xy}(z)$ and $\mu_x(z)$ routinely violate the cross-regime feasibility constraints (Eq. 5). Such estimates fail the compatibility requirements of a shared structural causal model, yielding internally inconsistent bounds. Therefore, instead of estimating these six quantities independently, we design an architecture that enforces these relationships by construction. The key idea is to treat the joint probabilities $p_{xy}(z)$ as *anchors*: primary outputs trained directly on observational data. The interventional probabilities $\mu_x(z)$ are then parameterised as constrained functions of these anchors.

Our architecture consists of a shared representation network $\phi_\theta : \mathcal{Z} \to \mathbb{R}^d$ that maps covariates to a learned embedding, followed by three output heads. The first head outputs logits over the four joint outcomes $(X, Y) \in \{0, 1\}^2$, which are passed through a softmax to produce the joint probabilities $p_{00}(z), p_{01}(z), p_{10}(z), p_{11}(z)$. The remaining two heads output unconstrained scalar values $\delta_0(z), \delta_1(z) \in \mathbb{R}$, which are transformed into the interventional probabilities $\mu_0(z), \mu_1(z)$ via a constraint-enforcing reparameterisation layer.

We enforce the Tian–Pearl constraints by reparameterising the interventional probabilities as convex combinations of their feasible bounds:

$$\mu_0(z) = p_{01}(z) + \big(1 - p_{00}(z) - p_{01}(z)\big)\,\sigma\big(\delta_0(z)\big), \tag{7}$$

$$\mu_1(z) = p_{11}(z) + \big(1 - p_{10}(z) - p_{11}(z)\big)\,\sigma\big(\delta_1(z)\big), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function. For each fixed z, the map $\delta_x(z) \mapsto \mu_x(z)$ is smooth and strictly increasing, since $\mu_x(z)$ is an affine function of $\sigma\big(\delta_x(z)\big)$ and σ is smooth and strictly increasing on $\mathbb{R}$. Writing the lower and upper feasible bounds for $\mu_x(z)$ as $\ell_x(z)$ and $u_x(z)$, the derivative $\partial\mu_x(z)/\partial\delta_x(z) = \big(u_x(z) - \ell_x(z)\big)\,\sigma\big(\delta_x(z)\big)\big(1 - \sigma\big(\delta_x(z)\big)\big)$ is strictly positive whenever the interval $[\ell_x(z), u_x(z)]$ is non-degenerate. Thus the reparameterisation is a smooth, order-preserving bijection from $\mathbb{R}$ onto the open feasible interval for $\mu_x(z)$. Gradient-based optimisation over $\delta_x(z)$ therefore explores the entire feasible region for $\mu_x(z)$ while preserving the constraints. Since $\sigma(\delta) \in (0, 1)$, each $\mu_x(z)$ lies within its feasible interval for any network weights, and the Tian–Pearl consistency constraints hold automatically. The joint-probability head is trained on observational data using a cross-entropy loss over the four-class outcome (X, Y). The auxiliary heads are trained on experimental data using binary cross-entropy losses on the derived interventional probabilities $\mu_0(z)$ and $\mu_1(z)$. When computing the experimental loss, we block gradients through the joint-probability branch used to construct the feasible bounds. The experimental loss updates the auxiliary heads and shared representation, while the joint-head-specific parameters receive direct gradients only from the observational loss.

3.2 Bias correction and uncertainty quantification

As mentioned previously, the max–min structure of PNS bounds induces systematic extremum bias even with consistent probability estimates.

Let $\psi^\star$ denote a scalar target parameter that is only partially identified. Now, suppose $\psi^\star$ is known to lie in the intersection $\psi^\star \in \bigcap_{v\in\mathcal{V}} \big[\,\psi_\ell(v),\, \psi_u(v)\,\big]$, where $\psi_\ell(v)$ and $\psi_u(v)$ are lower and upper bounding functions indexed by $v \in \mathcal{V}$. Equivalently, the identified set for $\psi^\star$ is the interval $[\psi_\ell^\star, \psi_u^\star]$ with $\psi_\ell^\star = \sup_{v\in\mathcal{V}} \psi_\ell(v)$, and $\psi_u^\star = \inf_{v\in\mathcal{V}} \psi_u(v)$.

Given estimators $\hat{\psi}_\ell(v)$ and $\hat{\psi}_u(v)$ with pointwise standard errors $s_\ell(v)$ and $s_u(v)$, a naive plug-in approach forms $\sup_v \hat{\psi}_\ell(v)$ and $\inf_v \hat{\psi}_u(v)$. The supremum tends to pick out values of v where estimation noise is positive ($\hat{\psi}_\ell(v) > \psi_\ell(v)$), producing an upward-biased estimate of the lower bound. To illustrate (Fig. 1), let the three estimated lower-bound terms be $\ell \in \{0.40, 0.28, 0.30\}$, together with the fixed term 0, yielding $\mathrm{PNS}_L(z) = 0.40$. A noisy estimator yielding $\hat{\ell} \in \{0.40, 0.68, 0.22\}$ produces a plug-in bound of $\hat{\mathrm{PNS}}_L(z) = \max\{0, \max \hat{\ell}\} = 0.68 > 0.40$. Because extremum operators naturally select the most severe estimation errors, this "winner's curse" systematically overestimates lower bounds and underestimates upper bounds, yielding spuriously tight intervals.

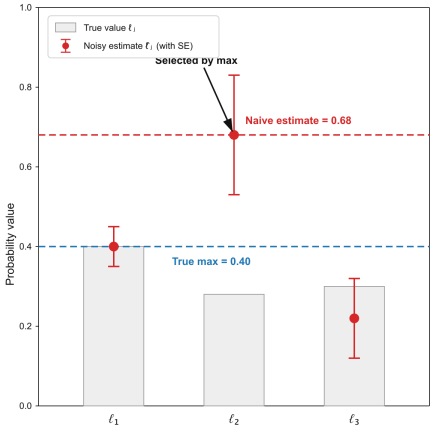


Figure 1: Overestimation bias among the three estimated lower-bound terms.

Conversely, the infimum tends to pick out values where noise is negative, producing a downward-biased estimate of the upper bound. The resulting bounds are therefore spuriously tight relative to the true identified set.

Therefore, in addition to ensuring structural causal compatibility, valid inference demands a framework that quantifies finite-sample uncertainty, and rigorously corrects for extremum bias. Towards that end, our inference procedure builds on the intersection bounds methodology of Chernozhukov et al. (2013b), which adjusts for this **maximisation bias** by penalising regions of high uncertainty. The uncertainty at each v is summarised by $s_\ell(v)$ and $s_u(v)$, which measure the sampling variability

of $\hat{\psi}_\ell(v)$ and $\hat{\psi}_u(v)$. The precision-corrected lower and upper bound estimators are defined as

$$\hat{\psi}_{\ell,\alpha} = \sup_{v\in\mathcal{V}} \big[\hat{\psi}_\ell(v) - \kappa_\alpha\, s_\ell(v)\big], \qquad \hat{\psi}_{u,\alpha} \quad = \inf_{v\in\mathcal{V}} \big[\hat{\psi}_u(v) + \kappa_\alpha\, s_u(v)\big], \tag{9}$$

where κ_α is a critical value chosen to ensure coverage at level $1-\alpha$. Intuitively, the penalty $\kappa_\alpha\, s_\ell(v)$ (resp. $\kappa_\alpha\, s_u(v)$) downweights regions of $\mathcal{V}$ where the estimator is imprecise, preventing the supremum (infimum) from concentrating on points with favourable but unreliable estimation error. The theoretical choice of κ_α is based on the standardised error process. Focusing on the upper bound (the lower bound is analogous), define

$$T_n(v) = \frac{\psi_u(v) - \hat{\psi}_u(v)}{s_u(v)}. \tag{10}$$

Under regularity conditions, Chernozhukov et al. show that $T_n(v)$ can be strongly approximated by a Gaussian process $T_n^\star(v)$ satisfying $\sup_{v\in\mathcal{V}} \big|T_n(v) - T_n^\star(v)\big| \xrightarrow{p} 0$. This approximation allows κ_α to be chosen as the $(1-\alpha)$-quantile of $\sup_{v\in\mathcal{V}} T_n^\star(v)$. In practice, one estimates the covariance structure of $T_n^\star$ from the data, simulates B draws from the approximating Gaussian process, computes the simulated suprema, and takes their empirical $(1-\alpha)$-quantile as κ_α. Moreover, Chernozhukov et al. (2013b) also prove that their precision-corrected estimator $\hat{\psi}_{u,\alpha}$ satisfies $\liminf_{n\to\infty} P\big(\psi_u^\star \le \hat{\psi}_{u,\alpha}\big) \ge 1-\alpha$. Under the regularity conditions stated in Appendix B.2.2, the resulting joint critical value yields asymptotically valid coverage; see Theorem B.1 and Chernozhukov et al. (2013b, Thm. 2).

Application to PNS bounds. The Tian–Pearl bounds on $\mathrm{PNS}(z)$ contain three estimated lower-bound terms, the fixed lower term 0, and four estimated upper-bound terms:

$$L(z) = \max\left\{0, \max_{j\in\{1,2,3\}} \ell_z(j)\right\}, \qquad U(z) = \min_{k\in\{1,\dots,4\}} u_z(k), \tag{11}$$

where each $\ell_z(j)$ and $u_z(k)$ is a known function of the observational and interventional probabilities at covariate value z. This is a finite-dimensional instance of the intersection bounds problem. The precision correction serves two roles in this application. As an estimator, it counters the systematic bias induced by the max and min operators selecting components with favourable noise. As an inference procedure, it provides simultaneous coverage of the bounding functionals which, combined with the algebraic validity of the Tian–Pearl inequalities at the true distribution, yields coverage of $\mathrm{PNS}(z)$.

Applying the precision correction requires standard errors $s_{\ell_z}(j)$ and $s_{u_z}(k)$ and a critical value κ_α derived from the estimator's sampling variability. Direct bootstrap estimation requires repeated model refitting, while classical influence-function approximations characterise the effect of each observation through $\psi(x) = -H^{-1}\nabla_\theta \ell(x,\hat{\theta})$, where H^{-1} propagates local parameter perturbations through the fitted estimator (Chernozhukov et al., 2013a). For a network with p parameters and n samples, explicitly constructing the dense Hessian costs approximately $O(np^2)$ and inverting it costs $O(p^3)$, with additional Hessian-vector operations required to evaluate uncertainty at target covariates. These costs are prohibitive when $p \gg 10^5$. We therefore use ENNs to generate jointly coherent function-level perturbations without explicit linearisation or Hessian inversion, providing a scalable surrogate for the finite-sample variability needed to estimate the standard errors and calibrate κ_α.

3.2.1 Epistemic Neural Networks for Precision-Corrected Inference

We use Epistemic Neural Networks (ENNs) (Osband et al., 2023) operationally as a frequentist uncertainty surrogate. Rather than interpreting ENN variability as a Bayesian posterior, we use epistemic draws to generate coherent perturbations of the fitted estimator, in the spirit of multiplier bootstrap resampling. This provides a scalable mechanism for approximating the joint variability required for inference in high-dimensional settings.

The ENN framework augments a base network with an epinet and an explicit epistemic index. This construction is agnostic to the internal architecture of the base network and allows us to sample from our constrained multi-head model without architectural changes. In our setting, ENNs are used to

generate stochastic perturbations of a fitted estimator that approximate finite-sample variability (Osband et al., 2018; 2023). These properties make ENNs a natural choice for our precision-corrected PNS bounds.

A standard neural network defines a function $f_\theta \colon \mathcal{Z} \to \mathbb{R}^k$. An ENN extends this to a map $f_\theta \colon \mathcal{Z} \times \mathcal{E} \to \mathbb{R}^k$, where $\mathcal{E}$ is the space of epistemic indices and P_ζ is a reference distribution on $\mathcal{E}$. For fixed z, the distribution of $f_\theta(z, \zeta)$ under $\zeta \sim P_\zeta$ represents the model's epistemic uncertainty about the prediction at z. When we consider a sequence of inputs $z_1, \dots, z_\tau$, the ENN defines a joint predictive distribution $\hat{P}_{1:\tau}(y_{1:\tau}) = \int_{\mathcal{E}} \prod_{t=1}^{\tau} P\big(y_t \mid f_\theta(z_t, \zeta)\big) dP_\zeta(\zeta)$. Dependence of the predictions across t through a shared ζ encodes reducible uncertainty that is coherent across inputs. In our setting the ENN outputs the six probability estimates that define the Tian–Pearl bounds. For each (z, ζ) we obtain $f_\theta(z, \zeta) = \big(p_{xy}(z; \theta, \zeta)_{x,y \in \{0,1\}}, \mu_0(z; \theta, \zeta), \mu_1(z; \theta, \zeta)\big)$, and the anchored architecture ensures that these probabilities satisfy the algebraic constraints for every ζ. For a fixed z we draw $\zeta_1, \dots, \zeta_M \sim P_\zeta$. For each m we compute the three estimated lower-bound terms $\ell_z^{(m)}(j)$ and four upper-bound terms $u_z^{(m)}(k)$ from the corresponding probabilities. The fixed lower term 0 is kept separate. We approximate the mean and standard deviation of each bound term under the epistemic distribution by Monte Carlo averages:

$$\begin{aligned} \bar{\ell}_z(j) &= \frac{1}{M} \sum_{m=1}^{M} \ell_z^{(m)}(j), \qquad s_{\ell_z}(j) = \sqrt{\frac{1}{M-1} \sum_{m=1}^{M} \big(\ell_z^{(m)}(j) - \bar{\ell}_z(j)\big)^2}, \\ \bar{u}_z(k) &= \frac{1}{M} \sum_{m=1}^{M} u_z^{(m)}(k), \quad s_{u_z}(k) = \sqrt{\frac{1}{M-1} \sum_{m=1}^{M} \big(u_z^{(m)}(k) - \bar{u}_z(k)\big)^2}, \end{aligned} \tag{12}$$

for $j \in \{1, 2, 3\}$ and $k \in \{1, \dots, 4\}$. The quantities $\bar{\ell}_z(j)$ and $\bar{u}_z(k)$ play the role of $\hat{\psi}_\ell(v)$ and $\hat{\psi}_u(v)$ in the generic intersection bounds notation. The quantities $s_{\ell_z}(j)$ and $s_{u_z}(k)$ are ENN-based analogues of the standard errors $s_\ell(v)$ and $s_u(v)$. Note that compared to the $O(np^2 + p^3)$ Hessian operations required by influence function computation, ENNs require only $O(M)$ forward passes, where M is the number of epistemic index samples at test time.

Critical value computation. The precision-corrected framework also requires a joint critical value that respects the dependence structure across the bound terms. In Chernozhukov et al. (2013b) this critical value is derived from a Gaussian approximation to the studentised estimation process. In our setting we cannot construct such a process directly, because we do not have an influence function representation. Instead we use the empirical distribution of ENN samples as a surrogate.

For each epistemic draw m we compute standardised deviations

$$\tilde{W}_z^{(m)}(j) = \frac{\ell_z^{(m)}(j) - \bar{\ell}_z(j)}{s_{\ell_z}(j)}, \quad \tilde{V}_z^{(m)}(k) = \frac{u_z^{(m)}(k) - \bar{u}_z(k)}{s_{u_z}(k)}. \tag{13}$$

We then form the maxima and minima

$$\tilde{W}_z^{L,(m)} = \max_{j \in \{1,2,3\}} \tilde{W}_z^{(m)}(j), \quad \tilde{W}_z^{U,(m)} = -\min_{k \in \{1,\dots,4\}} \tilde{V}_z^{(m)}(k). \tag{14}$$

The critical value κ_α^L is chosen as the empirical $(1-\alpha/2)$ quantile of the sample $\{\tilde{W}_z^{L,(m)}\}_{m=1}^M$. The critical value κ_α^U is defined analogously from $\{\tilde{W}_z^{U,(m)}\}_{m=1}^M$. Splitting the error probability equally between the lower and upper sides gives joint coverage of the resulting two-sided interval by the union bound. This construction mirrors the Gaussian process simulation step in Chernozhukov et al. (2013b), but uses the ENN-induced empirical distribution instead of an analytic Gaussian coupling.

Precision-corrected PNS bounds. The precision-corrected PNS bounds at covariate value z penalise terms with high epistemic variability. They are given by

$$\hat{L}_\alpha(z) = \max\Big\{0, \max_{j \in \{1,2,3\}} \big[\bar{\ell}_z(j) - \kappa_\alpha^L \, s_{\ell_z}(j)\big]\Big\}, \qquad \hat{U}_\alpha(z) \;\; = \min_{k \in \{1,\dots,4\}} \big[\bar{u}_z(k) + \kappa_\alpha^U \, s_{u_z}(k)\big]. \tag{15}$$

The interval $[\hat{L}_\alpha(z), \hat{U}_\alpha(z)]$ provides a precision-corrected estimate of the identified set for $\mathrm{PNS}(z)$ that adjusts for selection bias induced by the max and min operators. Complete pseudocode for the proposed procedure is provided in Appendix E, Algorithm 1.

Theoretical guarantee and scalable implementation. Under asymptotic linearity of the six atomic probability estimators, consistent plug-in influence functions, and multiplier-bootstrap validity, Theorem B.1 establishes asymptotic coverage of at least $1 - \alpha$ for the full Tian–Pearl identified set at each fixed z using the influence-function construction. Appendix B also provides sufficient conditions based on finite-dimensional M-estimation and, under stronger uniform regularity conditions, a simultaneous band over $\mathcal{Z}$ using a globally calibrated critical value. Causal EpiNet implements the same precision-correction rule using amortised draws from a constrained ENN (Osband et al., 2023), avoiding explicit influence-function and Hessian calculations. Each epistemic draw jointly determines all bound components while satisfying the model's probability constraints, supplying function-level perturbations for standard-error estimation and CLR calibration of κ_α. The coverage guarantee is established for the influence-function construction; the calibration and computational scalability of the ENN approximation are assessed through finite-sample coverage experiments and comparisons with full-network and last-layer multiplier bootstraps where computationally feasible (Section 4).

4 Experimental Evaluation

This section evaluates the finite-sample behaviour of machine-learning estimators for PNS bounds. The experiments address five questions: (1) do standard plug-in estimators violate the Tian–Pearl probability constraints, and does this degrade bound quality? (2) How much coverage improvement comes from precision correction, and how severe is extremum bias without it? (3) Does EpiNet uncertainty achieve coverage comparable to the full-network influence-function multiplier bootstrap, and at what computational cost? (4) How does bound quality change with the amount of experimental data? (5) In a large-scale application, what additional information do corrected PNS bounds provide beyond CATE-based treatment-effect heterogeneity?

4.1 Data-Generating Processes and Protocol

For the first four questions, we evaluate our method in two benchmark settings with fully specified SCMs. In both, the learner observes only a subset of covariates, inducing hidden confounding while preserving exact ground-truth quantities. The first uses the synthetic SCM of Li et al. (2022), with binary treatment X, binary outcome Y, and 20 binary covariates Z. The learner observes 15 covariates, while 5 remain hidden. The second uses the 200 real-valued covariates from ACIC (2019), augmented with 5 synthetic confounders. Treatments and potential outcomes are generated from an SCM over all 205 variables, while the learner observes only the original 200. In both settings, the known SCM provides exact $\mathrm{PNS}(z)$ values and sharp Tian–Pearl bounds, enabling direct evaluation of coverage and bound quality. Results use disjoint test splits and are averaged over 1,000 independent replicates. Further details appear in Appendices C and D.

Metrics. For each test point z_i, a method returns an interval $[\ell_i, u_i]$. Given the true $\mathrm{PNS}(z_i)$ and oracle identified set $[L(z_i), U(z_i)]$, we report: *constraint violation rate*, the fraction violating Eq. 5; *point coverage*, the fraction with $\mathrm{PNS}(z_i) \in [\ell_i, u_i]$; *identified-set coverage*, the fraction with $[L(z_i), U(z_i)] \subseteq [\ell_i, u_i]$; *bound width*, the mean $|u_i - \ell_i|$; and the *interval score* (Winkler, 1972), which penalises both width and miscoverage. Both coverage metrics target 95%. Full definitions are given in Appendix E.6.

4.2 Methods

We compare the proposed method with plug-in, constrained-network, and multiplier-bootstrap baselines. These comparisons separate the contributions of constraint enforcement, uncertainty estimation, and precision correction.

S-learner and T-learner plug-in. These are standard baselines that estimate the quantities entering the Tian–Pearl bounds independently, without any structural coordination. The S-learner fits a single model for $P(Y = 1 \mid X, Z)$ and obtains $\hat{\mu}_1(z)$ and $\hat{\mu}_0(z)$ by evaluating it at $X = 1$ and $X = 0$ respectively. The T-learner fits separate models on the treated and control arms to estimate $\mu_1(z)$ and $\mu_0(z)$ directly. In both cases, the joint observational probabilities

Table 1: Results on low-dimensional (top) and high-dimensional (bottom) datasets.

METHOD	% VALID (↑👍)	PT COV. (↑👍)	ID COV (↑👍)	WIDTH (↓👍)	INT. SCORE (↓👍)
LOW-DIMENSIONAL SYNTHETIC					
S-LEARNER	76.29 [75.81, 76.78]	74.52 [73.90, 75.13]	22.49 [22.21, 22.77]	0.1730 [0.1716, 0.1745]	0.7295 [0.7132, 0.7458]
T-LEARNER	74.63 [74.16, 75.11]	72.39 [71.82, 72.95]	27.46 [27.22, 27.70]	0.1782 [0.1772, 0.1793]	0.9874 [0.9695, 1.0054]
ANCHORED	100	82.34 [82.14, 82.54]	36.36 [36.11, 36.61]	0.2041 [0.2001, 0.2081]	0.8102 [0.8062, 0.8142]
MB FULL	100	98.50 [98.10, 98.90]	97.40 [97.0, 97.8]	0.3695 [0.3665, 0.3725]	0.6385 [0.6365, 0.6405]
MB LAST LYR.	100	88.10 [87.70, 88.30]	73.41 [73.38, 73.44]	0.2045 [0.2015, 0.2075]	0.5112 [0.5072, 0.5152]
ANCH. EPINET.	100	**98.19** [97.99, 98.59]	**95.59** [95.30, 95.80]	**0.3303** [0.3263, 0.3323]	**0.3471** [0.3451, 0.3501]
HIGH-DIMENSIONAL ACIC					
ANCHORED	100	85.70 [82.1, 87.95]	21.50 [20.4, 22.6]	0.1630 [0.160, 0.165]	0.7815 [0.7810, 0.7820]
MB LAST LYR.	100	87.50 [85.5, 89.5]	35.40 [33.4, 37.4]	0.2010 [0.180, 0.220]	0.7200 [0.7000, 0.7400]
ANCH. EPINET.	100	**97.30** [96.80, 97.70]	**86.00** [85.3, 86.7]	**0.3060** [0.301, 0.310]	**0.3630** [0.3520, 0.3740]

$p_{xy}(z) = P(X = x, Y = y \mid Z = z)$ are estimated by a separate multinomial model trained on the observational data. All estimates are substituted directly into the Tian–Pearl formulas with no constraint enforcement and no precision correction. **Anchored network (no CLR correction).** The anchored architecture introduced in Section 3.1 jointly estimates all six atomic probabilities $(\mu_1, \mu_0, p_{11}, p_{10}, p_{01}, p_{00})$ and enforces the Tian–Pearl consistency constraints by construction via the reparameterisation in Equations 7–8. Point estimates are plugged into the bound formulas to obtain interval endpoints. No precision correction is applied. This baseline isolates the contribution of constraint enforcement from that of the CLR correction: any improvement over the plug-in methods is due to the anchored architecture alone, and any remaining gap to the full method is attributable to the absence of precision correction. **Full-network multiplier bootstrap (MB Full).** This baseline uses the anchored network and computes influence functions with respect to all network parameters. Multiplier-bootstrap standard errors are combined with the same CLR precision correction used by the proposed method. We evaluate this baseline only in the low-dimensional setting because its computational and memory costs are prohibitive in the high-dimensional setting. **Last-layer multiplier bootstrap (MB Last Lyr.).** This baseline uses the same procedure but computes influence functions only for the final linear layer. This reduces the size of the required Hessian calculation and makes the method practical in the high-dimensional ACIC experiment. **Anchored EpiNet + CLR (full method).** This is the proposed method: the anchored architecture with an ENN for epistemic uncertainty quantification, combined with precision-corrected inference through the CLR construction. Additional details regarding the experimental setup can be found in Appendix E.

4.3 RESULTS

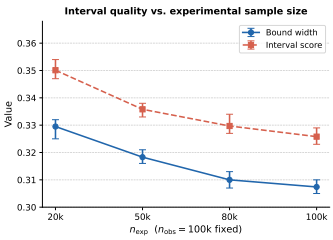


Figure 2: Interval quality versus n_{exp} ($n_{\text{obs}} = 100$k).

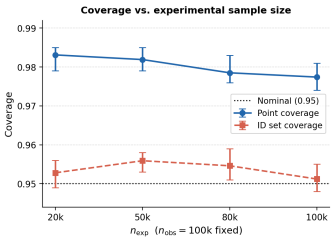


Figure 3: Coverage versus n_{exp} ($n_{\text{obs}} = 100$k).

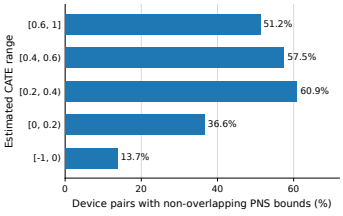


Figure 4: PNS-bound non-overlap within CATE ranges.

Table 1 reports results for the low-dimensional Li et al. benchmark ($n_{\text{obs}} = 100$k, $n_{\text{exp}} = 20$k) and the high-dimensional ACIC setting ($n_{\text{obs}} = 2$M, $n_{\text{exp}} = 1$M). Bracketed values are 95% confidence intervals over 1,000 replicates. Low-dimensional replicates use fresh samples from the exact SCM, while ACIC replicates subsample a fixed super-population of 3M observational and 3M experimental records. In the low-dimensional setting, S- and T-learner plug-ins attain point coverage of 0.745 and 0.724, with interval scores of 0.730 and 0.987. Their narrow intervals partly reflect infeasible probability configurations rather than improved sharpness. The anchored network enforces the constraints and raises coverage to 0.823, but its score remains 0.810. Causal EpiNet attains point

coverage of 0.982 $[0.980, 0.986]$, identified-set coverage of 0.956 $[0.953, 0.958]$, and the lowest interval score, 0.347. The full-network bootstrap gives similar point coverage (0.985), but with wider intervals (0.370 versus 0.330) and a higher score (0.639). The last-layer bootstrap undercovers at 0.881 $[0.877, 0.883]$.

In the high-dimensional ACIC setting, the full-network influence-function calculation is computationally intractable and is omitted. Causal EpiNet uniquely maintains valid point coverage at 0.973 $[0.968, 0.977]$ with a score of 0.363. Its identified-set coverage decreases to 0.860, consistent with residual error in estimating $p_{xy}(z)$ and $\mu_x(z)$. The uncorrected estimator and last-layer bootstrap undercover at 0.857 $[0.821, 0.880]$ and 0.875 $[0.855, 0.895]$, with scores of 0.782 and 0.720, respectively. The last-layer result is consistent with uncertainty from the learned representation not being captured when influence calculations are restricted to the final linear layer.

We evaluate Causal EpiNet across $(n_{\text{obs}}, n_{\text{exp}})$ regimes to assess behaviour under the practically common condition of scarce experimental data. Figures 2 and 3 show that as n_{exp} increases from 20k to 100k with $n_{\text{obs}} = 100$k fixed, identified-set coverage remains nominal throughout while bound width decreases monotonically. This confirms that ENN uncertainty estimates correctly track experimental data availability without sacrificing calibration. Full results, including balanced scaling up to 1M/1M, are reported in Appendix G. To assess computational efficiency, we record the training and inference wall-clock times for all evaluated methods. Because ENN inference uses batched forward evaluations rather than repeated model retraining or costly influence-function evaluations, it incurs substantially lower computational overhead than the bootstrap baselines. A comprehensive runtime breakdown is deferred to Appendix F.

Point coverage is necessarily at least as large as identified-set coverage: because $\text{PNS}(z) \in [L(z), U(z)]$, containment of $[L(z), U(z)]$ implies point coverage, but not conversely. The joint CLR construction targets endpoint containment. In the ACIC benchmark, identified-set coverage is 0.860 even though point coverage is 0.973, since the true PNS may remain inside the estimated interval when an oracle endpoint is missed.

Microsoft Product-Release Case Study To complement the preceding benchmarks, we apply Causal EpiNet to a large-scale product-release study containing 435,170 randomised and 445,286 observational devices (Young & Dillon, 2026). We compare corrected PNS bounds, estimated using both data sources, with cross-fitted CATE estimates obtained from the randomised sample. Within five prespecified CATE ranges, Figure 4 shows that 13.7% to 60.9% of device pairs have non-overlapping corrected PNS bounds, with median separations of 4.0 to 5.9 percentage points among non-overlapping pairs. This variation is compatible with similar CATE values because CATE measures benefit minus harm, whereas PNS isolates the probability of benefit. Complete data, estimation, and gap analyses are provided in Appendix H.

5 Discussion and Conclusion

We presented Causal EpiNets, a framework for finite-sample inference on individual-level causal effects through PNS bounds. Combining a constraint-preserving architecture, ENN uncertainty quantification, and CLR precision correction yields near-nominal coverage without requiring observational unconfoundedness or positivity. Instead, the method requires experimental and observational data from the same population and the stability assumption of Section 2. It also extends directly to the probability of harm $P(Y_{x'} = 1, Y_x = 0 \mid Z = z)$, enabling separate analysis of benefit and harm in the binary setting. In the Microsoft case study, devices within the same CATE range can have non-overlapping corrected PNS bounds separated by several percentage points, showing that similar estimated net effects can mask differences in the estimated probability of benefit.

Limitations and future work. The current framework is restricted to binary treatments and outcomes. Although identification results exist for multi-valued and continuous treatments (Kawakami et al., 2024; Li & Pearl, 2024), scalable estimation and inference remain open. Extension to unstructured covariates such as text and images will also require understanding how learned representations interact with structural constraints and ENN uncertainty.

Public datasets containing both experimental and observational samples remain scarce, making synthetic and semi-synthetic benchmarks necessary for evaluating coverage against ground-truth PNS. Key directions include extending the framework beyond binary settings, incorporating modern representation learning, and developing smoothing-based inference for conditional PNS. The latter requires orthogonal scores for the six jointly constrained probabilities and joint treatment of the lower and upper envelopes, providing an alternative to direct precision correction of the Tian–Pearl max–min bounds.

REFERENCES

ACIC. Acic 2019 data challenge datasets. https://sites.google.com/view/acic2019datachallenge/data-challenge, 2019.

Ahmed M. Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, 1993.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *Conference on Learning Theory (COLT)*, 2020.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Weilin Chen, Ruichu Cai, Zeqin Yang, Jie Qiao, Yuguang Yan, Zijian Li, and Zhifeng Hao. Doubly robust causal effect estimation under networked interference via targeted learning. *International Conference on Machine Learning (ICML)*, 2024.

Xiaohong Chen and Halbert L. White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999. doi: 10.1109/18.749011.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6): 2786–2819, 2013a. doi: 10.1214/13-AOS1161.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737, 2013b. doi: 10.3982/ECTA8718.

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. *International Conference on Learning Representations (ICLR)*, 2024.

Nathan Kallus. What's the harm? sharp bounds on the fraction negatively affected by treatment. *arXiv preprint arXiv:2205.10327*, 2022. doi: 10.48550/arXiv.2205.10327. URL https://arxiv.org/abs/2205.10327.

Yuta Kawakami, Manabu Kuroki, and Jin Tian. Probabilities of causation for continuous and vector variables. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1901–1921. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/kawakami24a.html.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1885–1894, 2017.

Hui Lan and Vasilis Syrgkanis. Causal Q-aggregation for CATE model selection. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4366–4374. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/lan24a.html.

Ang Li and Judea Pearl. Probabilities of causation with nonbinary treatment and effect. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, volume 38 of *AAAI Technical Tracks*, pp. 20465–20472. Association for the Advancement of Artificial Intelligence (AAAI), Mar 2024. doi: 10.1609/aaai.v38i18.30030. URL https://ojs.aaai.org/index.php/AAAI/article/view/30030.

Ang Li, Ruirui Mao, and Judea Pearl. Probabilities of causation: Adequate size of experimental and observational samples. *arXiv preprint arXiv:2210.05027*, 2022. doi: 10.48550/arXiv.2210.05027.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *Conference on Learning Theory (COLT)*, 2018.

Jiannan Lu, Peng Ding, and Tirthankar Dasgupta. Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43(5):540–567, 2018.

James Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 735–742, 2010.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2nd edition, 1989. doi: 10.1007/978-1-4899-3242-6.

Scott Mueller and Judea Pearl. Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1), 2023. doi: 10.1515/jci-2022-0050.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden (eds.), *Handbook of Econometrics*, volume 4, chapter 36, pp. 2111–2245. Elsevier, 1994. doi: 10.1016/S1573-4412(05)80005-4.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 2023. URL https://arxiv.org/abs/2107.08924.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1–2):93–149, 1999. doi: 10.1023/A:1005233831499.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Md Musfiqur Rahman, Muhammad Ashad Kabir Farhad, Karthik Venkateswaran, and Laya Pradhan. UA-DCM: Uncertainty-aware causal decision making via effect bound decomposition. *arXiv preprint arXiv:2601.22736*, 2026.

Vira Semenova. Debiased machine learning of set-identified linear models. *Journal of Econometrics*, 235:1725–1746, 2023.

Xiaoxi Shen, Chao Jiang, Lyudmila Sakhanenko, and Qing Lu. Asymptotic properties of neural network sieve estimators. *Journal of Nonparametric Statistics*, 35(4):839–868, 2023. doi: 10.1080/10485252.2023.2209218.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1–4):287–313, 2000. doi: 10.1023/A:1018912507879. https://ftp.cs.ucla.edu/pub/stat_ser/r271-A.pdf.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9780511802256.

A. W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996. doi: 10.1007/978-1-4757-2545-2.

Shuai Wang and Ang Li. Estimating probabilities of causation with machine learning models. *arXiv preprint arXiv:2502.08858*, 2025.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. doi: 10.2307/1912526.

Halbert White. Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, 84(408):1003–1013, 1989. doi: 10.1080/01621459.1989.10478865.

Justin Whitehouse, Qizhao Chen, Morgane Austern, and Vasilis Syrgkanis. Inference on optimal policy values and other irregular functionals via softmax smoothing. *arXiv preprint arXiv:2507.11780*, 2026.

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10823–10836, 2021.

Justin Young and Eleanor Wiske Dillon. Reevaluating causal estimation methods with data from a product release. *arXiv preprint arXiv:2601.11845*, 2026. doi: 10.48550/arXiv.2601.11845. URL https://arxiv.org/abs/2601.11845.

AI USAGE DISCLOSURE

In this work, we used generative AI tools (Large Language Models) to assist with drafting and editing the manuscript, specifically for shortening prose, correcting typographical errors, and generating LaTeX formatting. Additionally, we used LLMs as a programming aid to help comprehend public repository code and to check our own code for errors or inefficiencies. We have not used generative AI tools to formulate original scientific claims, prove theorems, or generate our experimental data.

We have reviewed all AI-assisted work. Specifically, all LLM-edited text was manually verified by the authors to ensure the preservation of our original scientific meaning and intent. Furthermore, any code checked or explained by the LLMs was manually reviewed, and our final scripts were thoroughly tested for correctness.

ETHICS STATEMENT

This work introduces a theoretical estimation framework for bounding the probability of benefit. While our primary contribution is methodological and lacks direct ethical risks, such as the use of human subjects or sensitive data, we recognise that causal inference tools are frequently deployed in high-stakes domains like personalised medicine and public policy. In practice, the validity of these bounds depends on core causal assumptions, including valid experimental assignment and stability across the observational and experimental regimes. We encourage practitioners to rigorously verify these assumptions before deploying such estimates in the real world, as incorrect causal inferences can lead to harmful decisions. Ultimately, we anticipate that more rigorous estimation frameworks will support safer and more equitable decision-making.